Dealing with large, distributed scientific datasets on exascale computers

Michela Taufer

University of Delaware

1. Exascale challenges in large dataset analysis

Exascale platforms are expected to perform large-scale, computationally expensive simulations at a rate never seen before. However, this new capability of performing longer simulations will present new challenges for scientists who have to deal with the analysis, sorting, and selection of scientifically meaningful results from the massive amounts of data collected. To make the situation even more challenging, data can be located across distributed nodes of the exascale platform. Techniques are needed to analyze and characterize the large-scale, distributed scientific datasets. However, state-of-the-art techniques such as clustering require comparing data with each other many times, in an iterative process. When speaking about massive datasets distributed across a large number of nodes, as they could be in exascale systems, even a small number of comparisons have a great impact on the efficiency of the analysis algorithm.

2. Envisioned efforts and applicability

To address this incumbent problem, these key research questions are of great priority:

- Can we intrinsically encode data in large and distributed datasets, so that we can more effectively capture their semantic properties, while syntactically redefining the data structures for the sake of their analysis accuracy and scalability?
- Can we redesign linear-in-complexity algorithms for analyzing the encoded datasets, so that we can extract relevant scientific conclusions from the complete dataset?
- Can we integrate these algorithms into emerging distributed paradigms such as the MapReduce paradigm and implement them into MapReduce middleware packages, such as Hadoop and MapReduce-MPI?
- Can we measure scalable performance and accurate results for relevant scientific datasets, e.g., high-throughput protein-ligand docking datasets and geographical datasets?

We believe that these questions can be effectively answered for petascale systems today and exascale systems tomorrow in diverse scientific fields dealing with diverse datasets. The datasets can range from 3D molecular structures generated in e.g., high-throughput molecular dynamics (MD) simulations, such as the millions of peptide conformations in a folding simulation, or ligand conformations docked into a protein pocket in protein-ligand docking simulations, to geographical data structures for e.g., indexing methods for GIS Geographical Information Systems. Our preliminary results of a novel approach we present in this white paper, which were previously published in [1, 2], support our claim.

3. Proposed approach, its uniqueness, and maturity

In our preliminary work, we first provided positive answers to the questions above for a large dataset of protein-ligand conformations in drug design and studied the effectiveness of our approach, to achieve both accuracy and scalability on petascale architectures. A crucial step in the protein-ligand docking process is the accurate prediction of the binding geometry of a ligand from an ensemble of docked ligand conformations, which requires the evaluation of numerous possible predicted protein-ligand geometries in the order of billions of conformations [3, 4]. In evaluating the ensemble of possible protein-ligand binding geometries, scientists typically rely on the traditional scoring approach, based on a molecular mechanics energy function. We observed how conformations scoring minimum energy over very large datasets produced by billions of docking attempts might be significantly different from the experimentally observed conformation. On the other hand, if we compare the geometry of each ligand conformation in

the large dataset, we can find several conformations that are close to the near-native conformation, but would not necessary score the lowest energy [5]. Powerful hierarchical clustering methods are able to select near-native poses with higher accuracy, but unfortunately, they scale poorly for very large datasets [5]. However, in [1], we first showed how we can encode the geometry of each three-dimensional (3D) ligand conformation in a dataset of conformations into a single 3D point in space in a decoupled and completely distributed way, by projecting the 3D atoms of each conformation on each of the three 2D planes (x,y), (y,z), and (z,x) and computing the best-fit linear regression line of the 2D points. Secondly, we presented how we can redesign the clustering algorithm of the 3D ligand conformations into a density search that scales linearly when we use an octree representation of the space. Thirdly, we showed how we can integrate the density search into the MapReduce paradigm and implement it into Hadoop[1], one of the most utilized MapReduce implementations. Last but not least, we measured scalable performance and accurate results for our approach on the Gordon supercomputer at the San Diego Supercomputer Center (SDSC), on which, when using large datasets of billions of ligands, as the number of cores in the computer system increases, the execution time of our algorithm decreases linearly from 5 hours to 20 minutes, showing a performance improvement of nearly two orders of magnitude.

The proposed work is far from being completed. Research directions to be pursued by the investigator and her group include: (1) encoding a more diverse set of properties than the geometries of the molecules in our encoding method, e.g., to include not only the geometry but also the location of the ligand into the protein pocket or the presence of specific atoms or charges in a molecule; (2) extending our encoding approach to a broader type of dataset produced by MD simulations, e.g., protein-protein binding geometries, as well as protein conformations, for protein structure prediction and protein folding; (3) extending our encoding approach to a completely different research field and dataset, e.g., large geographical datasets; (4) further searching for alternate algorithms for data comparison and for their efficient integration into distributed frameworks; and (5) studying accuracy and performance of the overall approach on larger and larger computing platforms, while waiting for exascale platforms to become a reality for day-to-day scientific simulations. When exascale systems arrive on the scene, we will be ready to analyze scientists' data!

4. References

[1] T. Estrada, B. Zhang, P. Cicotti, R. Armen, and M. Taufer: A Scalable and Accurate Method for Classifying Protein-Ligand Binding Geometries using a MapReduce Approach. *Computers in Biology and Medicine,* 2012. (in press)

[2] T. Estrada, B. Zhang, P. Cicotti, R. Armen, and M. Taufer: Reengineering high-throughput molecular datasets for scalable clustering using MapReduce. In *Proceedings of the 14th IEEE International Conference on High Performance Computing and Communications (HPCC-2012)*, June 2012, Liverpool, England, UK.

[3] M. Taufer, R.S. Armen, J. Chen, P.J. Teller, and C.L. Brooks III: Computational Multi-Scale Modeling in Protein-Ligand Docking. *IEEE Engineering in Medicine and Biology Magazine*, 28(2): 58 – 69, 2009.

[4] O. Rahaman, T. Estrada, D. Doren, M. Taufer, C. L. Brooks III, R.S. Armen: Evaluation of Several Two-Step Scoring Functions Based on Linear Interaction Energy, Effective Ligand Size, and Empirical Pair Potentials for Prediction of Protein-Ligand Binding Geometry and Free Energy. *J. Chemical Information and Modeling*, 51(9): 2047 – 65, 2011.

[5] T. Estrada, R. Armen, and M. Taufer: Automatic Selection of Near-Native Protein-Ligand Conformations using a Hierarchical Clustering and Volunteer Computing. In *Proceedings of the ACM International Conference on Bioinformatics and Computational Biology (BCB),* August 2010, New York, USA.

[1] We recently completed the implementation of the same algorithm for MapReduce-MPI